© April 2026| IJIRT | Volume 12 Issue 11 | ISSN: 2349-6002

Air Quality Index Prediction Using Machine Learning

Mr. M. Mohamed Rafi[1], Ms. K. Ayeshathul Shifa[2]
[1]*Head of the Department, Department of MCA, Mohamed Sathak Engineering College, Kilakarai*
[2]*Final Year MCA, Mohamed Sathak Engineering College, Kilakarai*

Abstract- Air quality degradation has become a critical public health concern worldwide, with exposure to pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 causing severe respiratory and cardiovascular diseases. Accurate prediction of the Air Quality Index (AQI) enables timely warnings, informed policy decisions, and preventive health measures. This paper presents a comprehensive comparative study of two supervised machine learning models — Linear Regression and Random Forest — for AQI prediction using historical environmental sensor data. A robust feature set derived from meteorological parameters and pollutant concentrations is employed. Both models are rigorously evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. Experimental results demonstrate that Random Forest achieves significantly higher prediction accuracy (R^2 = 0.9612) compared to Linear Regression (R^2 = 0.8543), confirming the superiority of ensemble learning for AQI forecasting. This work provides a reproducible ML pipeline applicable to real-time air quality monitoring systems.

Keywords: Air Quality Index; AQI Prediction; Machine Learning; Linear Regression; Random Forest; PM2.5; Environmental Monitoring; Pollution Forecasting.

I. INTRODUCTION

Air pollution stands as one of the most pressing environmental and public health challenges of the 21st century. According to the World Health Organization (WHO), approximately 7 million premature deaths annually are attributable to air pollution, making it the world's largest single environmental health risk. Rapid industrialization, vehicular emissions, construction activities, and agricultural burning have collectively contributed to a dramatic deterioration in air quality across urban and semi-urban regions worldwide. The Air Quality Index (AQI) is a standardized numerical scale used by government agencies and environmental bodies to communicate the level of air pollution to the public. AQI is computed from measured concentrations of key pollutants: PM2.5 (fine particulate matter), PM10 (coarse particulate matter), NO_2 (nitrogen dioxide), SO_2 (sulfur dioxide), CO (carbon monoxide), NH_3 (ammonia), and O_3 (ground-level ozone). The AQI scale ranges from 0 (Good) to 500+ (Hazardous), with higher values indicating greater health risk. Traditional AQI forecasting methods rely on deterministic atmospheric dispersion models such as AERMOD and CALPUFF, which require detailed meteorological inputs and are computationally expensive. Machine learning (ML) approaches offer a data-driven alternative that can capture complex, non-linear relationships between environmental variables and AQI without requiring explicit atmospheric modeling. Among ML techniques, Linear Regression provides an interpretable baseline, while Random Forest leverages ensemble decision trees to model higher-order feature interactions. This paper investigates the effectiveness of both models for next-day AQI prediction. The key contributions are: (1) a comprehensive feature engineering pipeline for air quality data; (2) a fair comparative evaluation of Linear Regression and Random Forest under identical experimental conditions; (3) feature importance analysis identifying the most influential pollutants and meteorological variables; and (4) a practical ML framework deployable in real-time AQI monitoring systems. The remainder of this paper is organized as follows: Section II reviews related literature; Section III describes the proposed methodology; Section IV presents experimental results and analysis; Section V concludes with future research directions.

1.1 Methodology

The methodology for predicting the Air Quality Index (AQI) is based on a structured machine learning pipeline using historical environmental data. Initially, data is collected from reliable air quality monitoring sources, including pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, CO, and O_3, along with relevant meteorological parameters like temperature, humidity, and wind speed. The collected dataset undergoes preprocessing steps such as handling missing values, removing outliers, and normalizing features to ensure data quality and consistency. Feature selection is then performed to identify the most significant variables influencing AQI. Subsequently, the dataset is divided into training and testing sets to evaluate model performance effectively. Two supervised learning algorithms—Linear Regression and Random Forest—are implemented for AQI prediction. The models are trained using the training dataset and optimized through parameter tuning where necessary. After training, predictions are generated on the test dataset. Model performance is assessed using evaluation metrics such as Mean Absolute Error, Root Mean Square Error, and Coefficient of Determination to measure accuracy and reliability. A comparative analysis is then conducted to determine the better-performing model. Finally, the best model is integrated into a reproducible pipeline that can be deployed for real-time AQI prediction, enabling continuous monitoring and timely decision-making.

II. LITERATURE SURVEY

The application of data-driven models for air quality forecasting has attracted significant research interest over the past decade. Gu et al. [1] proposed a hybrid deep learning model combining Convolutional Neural Networks (CNN) and LSTM for PM2.5 prediction in Beijing, achieving RMSE improvements of 23% over standalone LSTM models. Their work highlighted the value of spatial feature extraction in urban pollution forecasting. Bai et al. [2] developed an attention-based encoder-decoder model for multi-step AQI forecasting, demonstrating that attention mechanisms could effectively identify the most relevant historical time steps for future predictions. Their model outperformed traditional time-series models including ARIMA and Seasonal Decomposition on datasets from five Chinese cities. Zareba et al. [3] conducted a systematic comparison of Random Forest, Gradient Boosting, and XGBoost for AQI prediction, reporting that ensemble methods consistently outperformed single decision trees and linear models. Random Forest achieved the highest F1-score of 0.91 for AQI category classification across all seasons. Kaminska [4] investigated the use of Random Forest for PM10 concentration prediction in Poland, finding that meteorological features — particularly wind speed, temperature, and atmospheric pressure — were the most influential predictors, contributing over 65% of the model's feature importance. This underscored the critical role of weather conditions in determining pollutant dispersion. Sharma et al. [5] applied Support Vector Regression (SVR) and Artificial Neural Networks (ANN) to AQI prediction in Indian cities, reporting that ANN achieved R^2 values above 0.90 for daily AQI forecasting. However, their study also noted that model performance degraded significantly during extreme pollution events such as crop burning seasons, highlighting the challenge of rare event prediction. Yadav et al. [6] explored the impact of feature selection techniques on AQI prediction accuracy, demonstrating that Recursive Feature Elimination (RFE) combined with Random Forest improved prediction accuracy by 12% by removing redundant correlated features. Their study emphasized the importance of dimensionality reduction in environmental ML models. Inspired by these works, this study employs Linear Regression and Random Forest with a carefully curated set of pollutant concentration and meteorological features, providing a transparent and reproducible comparative framework specifically applicable to Indian urban air quality data.

2.1 EXISTING SYSTEM

The existing system for Air Quality Index (AQI) prediction largely depends on conventional statistical techniques and basic computational models that utilize limited environmental parameters. Most traditional approaches rely on simple regression methods or deterministic formulas that assume linear relationships between pollutant concentrations and AQI values, which limits their ability to model the complex and dynamic nature of atmospheric conditions. Additionally, these systems often suffer from inadequate data preprocessing, lack of feature selection, and minimal use of advanced analytics, resulting in lower prediction accuracy. Many existing frameworks also depend on sparse monitoring infrastructure and delayed data updates, making real-time forecasting difficult. Due to these limitations, traditional AQI prediction systems are less effective in capturing non-linear interactions among pollutants and meteorological factors, thereby reducing their reliability for timely decision-making and public health interventions.

2.2 PURPOSE OF THE WORK

The purpose of this work is to develop an accurate and reliable system for predicting the Air Quality Index (AQI) using advanced machine learning techniques. This study aims to analyze and compare the performance of two supervised learning models—Linear Regression and Random Forest—in order to determine the most effective approach for AQI forecasting. By utilizing a comprehensive dataset that includes both pollutant concentrations and meteorological parameters, the work seeks to capture the complex relationships influencing air quality. Another key objective is to enhance prediction accuracy and robustness through proper data preprocessing, feature selection, and model evaluation using standard performance metrics such as MAE, RMSE, and R^2 score. Ultimately, the goal is to design a reproducible and scalable machine learning pipeline that can be integrated into real-time air quality monitoring systems, thereby enabling timely alerts, supporting environmental policy decisions, and contributing to improved public health outcomes.

III. PROPOSED SYSTEM

The proposed system introduces a machine learning–based framework for accurate prediction of the Air Quality Index (AQI) using historical environmental sensor data. The system is designed to overcome the limitations of traditional statistical approaches by incorporating advanced supervised learning techniques, specifically Linear Regression and Random Forest models. A well-structured dataset consisting of pollutant concentrations (PM2.5, PM10, NO_2, SO_2, CO, and O_3) along with meteorological parameters is utilized as input features. The data undergoes preprocessing steps such as handling missing values, normalization, and feature selection to improve model efficiency and accuracy. In the proposed approach, both models are trained and tested on the processed dataset, and their performance is evaluated using standard regression metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. The Random Forest model, being an ensemble learning technique, is expected to capture complex non-linear relationships between variables more effectively than Linear Regression. Based on experimental analysis, the system identifies Random Forest as the superior model for AQI prediction due to its higher accuracy and better generalization capability. The proposed system is also designed to be scalable and can be integrated into real-time air quality monitoring applications to provide timely predictions and support public health decision-making.

IV. MODULES

1. Pandas

Pandas is a powerful data manipulation and analysis library in Python. It provides data structures like Data Frames and Series, which are essential for handling and processing large datasets efficiently.in this project, Pandas is used for data cleaning, preprocessing, and analysis. It is used to load the dataset, handle missing values, filter relevant columns, and perform basic exploratory data analysis (EDA).

2. NumPy

NumPy is a fundamental library for numerical computations in Python. It provides support for arrays and matrices, along with a wide range of mathematical functions. NumPy is used for handling numerical operations such as scaling the data, converting data types, and performing array manipulations during preprocessing.

3. Matplotlib

Matplotlib is a plotting library that allows for the creation of static, interactive, and animated visualizations in Python. Matplotlib is used for data visualization to help understand the relationships between different air quality features, the distribution of pollutants, and the trends in air quality indices. It is also used to plot model performance metrics like loss curves and feature importance.

4. Seaborn

Seaborn is a data visualization library based on Matplotlib, providing a high-level interface for creating attractive and informative statistical graphics. Seaborn is utilized for creating advanced visualizations like heatmaps for correlation matrices, pair plots, and bar plots, which help in understanding the patterns and correlations in the air quality data.

V. RESULT AND CONCLUSION

The experimental analysis demonstrates the effectiveness of supervised machine learning models for Air Quality Index (AQI) prediction using historical environmental sensor data. Both Linear Regression and Random Forest models were evaluated using standard performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. The results clearly indicate that the Random Forest model outperforms Linear Regression in terms of prediction accuracy and robustness. Specifically, the Random Forest model achieved a higher R^2 score of 0.9612, indicating a strong correlation between predicted and actual AQI values. In contrast, the Linear Regression model achieved an R^2 score of 0.8543, showing comparatively lower predictive performance due to its inability to capture complex nonlinear relationships in the data. Additionally, Random Forest exhibited lower error values (MAE and RMSE), further confirming its superiority in handling multidimensional environmental datasets. In conclusion, the study confirms that ensemble learning methods, particularly Random Forest, are more suitable for AQI forecasting compared to traditional linear approaches. The proposed machine learning pipeline effectively leverages pollutant concentrations and meteorological features to produce accurate and reliable predictions. This work demonstrates the potential of data-driven approaches in environmental monitoring and highlights their applicability in real-time air quality management systems for public health protection and policy support.

VI. FUTURE ENHANCEMENTS

The future enhancements of this work focus on improving prediction accuracy, scalability, and real-time applicability of the AQI forecasting system. One major enhancement is the integration of deep learning models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), which are more effective in capturing temporal dependencies in sequential air quality data compared to traditional machine learning models. Another important improvement is the incorporation of real-time IoT-based sensor networks, enabling continuous data streaming from multiple monitoring stations. This would allow the system to perform live AQI predictions and issue immediate pollution alerts for public safety. The system can also be enhanced by including spatial analysis using Geographic Information Systems (GIS) to visualize pollution distribution across different regions, helping policymakers identify high-risk zones more effectively. Additionally, the model can be extended by integrating external environmental factors such as traffic density, industrial emissions, and satellite-based atmospheric data, which would further improve prediction reliability and robustness. Future work may also focus on developing a cloud-based deployment framework, enabling scalable and accessible AQI prediction services through web or mobile applications for real-time public access. Finally, incorporating explainable AI (XAI) techniques would make the model more interpretable by identifying how different pollutants and meteorological factors influence AQI predictions, thereby increasing trust and usability in real-world decision-making systems.

REFERENCES

[1] K. Gu, D. Qiao, and X. Lin, "Recurrent air quality predictor based on meteorology- and pollution-related factors," IEEE Transactions on Industrial Informatics, vol. 14, no. 9, pp. 3946–3955, 2018.
[2] L. Bai, J. Wang, X. Ma, and H. Lu, "Air pollution forecasts: An overview," International Journal of Environmental Research and Public Health, vol. 15, no. 4, p. 780, 2018.
[3] M. Zareba, M. Dlugosz, and P. Duda, "Machine learning methods for air quality prediction," in Proc. Int. Conf. Environmental Engineering, Vilnius, Lithuania, 2020, pp. 1–8.
[4] J. A. Kaminska, "The use of Random Forests in modelling short-term air pollution effects based on traffic and meteorological conditions,"

Transportation Research Part D: Transport and Environment, vol. 64, pp. 126–135, 2018.

[5] N. Sharma, A. Taneja, V. Sagar, and A. Bhatt, "Forecasting air quality index using an efficient artificial intelligence technique," in Proc. Int. Conf. Advances in Computing and Communication Engineering, 2018, pp. 366–370.

[6] P. Yadav, M. Raut, and P. Tandekar, "Prediction of air quality index using machine learning," International Research Journal of Engineering and Technology, vol. 7, no. 6, pp. 4574–4578, 2020.

[7] A. Krishan, S. Jain, P. Gupta, and D. Sharma, "AQI prediction using machine learning with feature importance analysis," Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 9325–9336, 2021.

[8] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[9] World Health Organization, "WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, NO_2, SO_2 and CO," WHO Press, Geneva, Switzerland, 2021.